UCF
CENTER FOR RESEARCH IN COMPUTER VISION

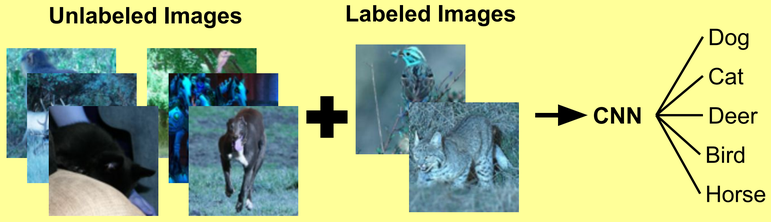
Semi-Supervised Training of CNNs

Kevin Duarte, Yang Zhang, and Boqing Gong
Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816

Problem & Motivation

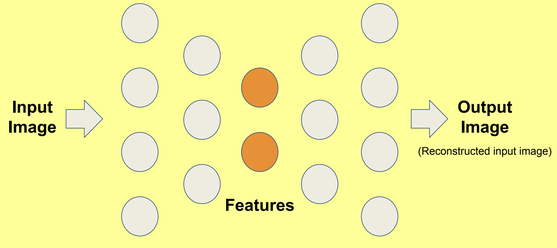
Problem: Building a convolutional neural network for image classification requires a large amount of labeled images to train. However, manually labeling images is a tedious and time-consuming process.

Semi-Supervised Training: Instead of using only labeled images, semi-supervised training takes advantage of unlabeled images to train an image classifier. We use a large set of unlabeled images and a small set of labeled image to train the CNN.



Approach

Step 1: Train an auto-encoder to compress and decompress unlabeled images in order to learn image features



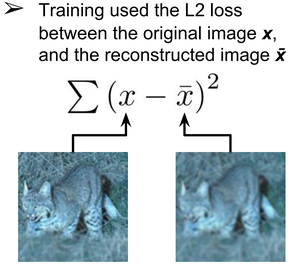
Step 2: Use learned features and a small amount of labeled images to train a CNN to classify images

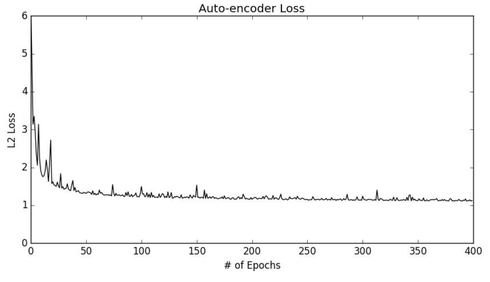
Experiments

Experiment Setup:

- Trained auto-encoder on 50,000 images from the cifar10 dataset
  - Compresses 3 channel, 32x32 images to 3 channel, 16x16 images
  - Auto-encoder structure: Input - (256)3c - (128)3c - 2p - (64)3c - (3)3c - (64)3c - 2up - (128)3c - (256)3c - (3)3c - Output
    - Input and output are 32x32 RGB images, (128)3c denotes a convolutional layer with 128 feature maps and a filter of size 3x3, 2p means 2x2 max-pooling, and 2up denotes a 2x2 upscaling layer
- Training used the L2 loss between the original image $\boldsymbol{x}$, and the reconstructed image $\bar{\boldsymbol{x}}$

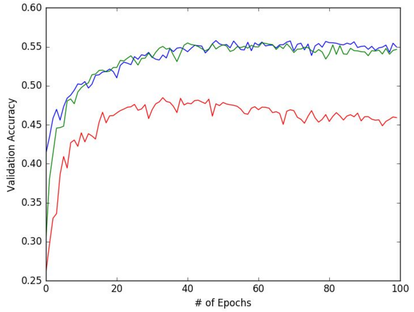
$$\sum (x - \bar{x})^2$$





Experiment Results:

- Trained 3 CNNs on the cifar10 dataset
  - CNN Structure: Input - (256)3c - (128)3c - 2p - (64)3c - (3)3c - 128fc - 10fc
  - Input is a 32x32 RGB image, 128fc denote a fully connected layer with 128 nodes, and the output is the image's class

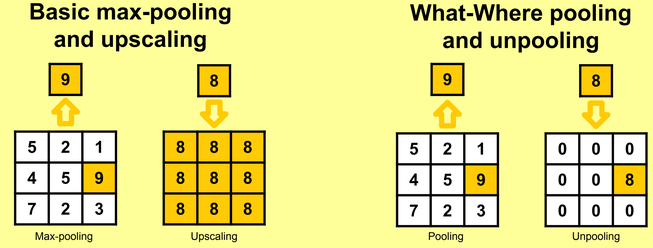


CNN	Maximum Accuracy
Semi-supervised, trained with 10,000 images	55.51%
Randomly initialized, trained with 10,000 images	48.47%
Randomly initialized, trained with 50,000 images	55.79%

- Using the same amount of labeled data, semi-supervised training greatly improved classification accuracy
- Semi-supervised training accuracy was equivalent to fully supervised training that used all 50,000 labeled images

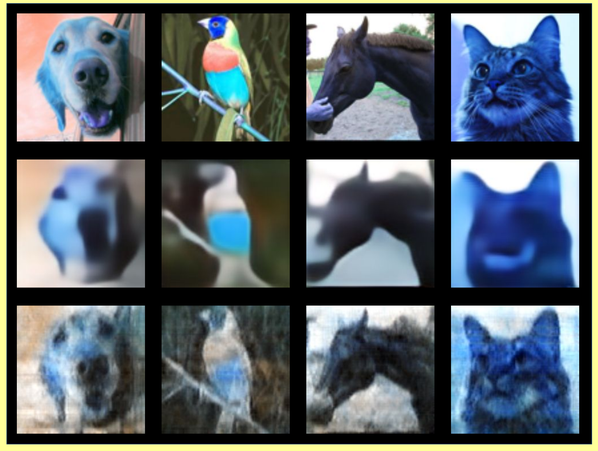
Stacked What-Where Auto-encoder

- A stacked what-where auto-encoder (SWWAE) replaces the max-pooling and upscaling layers used in a basic auto-encoder with what-where pooling/unpooling
  - SWWAEs have been shown to greatly improve semi-supervised training [1]



- We modified a VGG16 network and created a deeper auto-encoder
  - We trained a basic auto-encoder and a SWWAE with this deeper structure

SWWAE Results



[1] Junbo Zhao, Michael Mathieu, Ross Goroshin, and Yann Lecun. Stacked what-where auto-encoders. 2015. arXiv:1506.02351